DOCUMENT RESUME

ED 131 093 TM 005 761

AUTHOR Frederiksen, Norman
TITLE How to Tell if a Test Measures the Same Thing in Different Cultures.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RM-76-7
PUB DATE Aug 76
NOTE 12p.; Paper presented at the Congress of the International Association of Cross-Cultural Psychology (3rd, Tilburg, The Netherlands, July 13, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Covariance; *Culture Free Tests; Factor Analysis; Item Analysis; *Statistical Analysis; *Testing Problems; *Test Validity

ABSTRACT

A number of different ways of ascertaining whether or not a test measures the same thing in different cultures are examined. Methods range from some that are obvious and simple to those requiring statistical and psychological sophistication. Simpler methods include such things as having candidates "think aloud" and interviewing them about how they solved the problem, and techniques such as using pantomime or moving pictures to give instructions. Another approach is to make the tests different in such a way that they measure the same construct--so that they are functionally equivalent. The variety of approaches that require statistical methods include analysis of covariance, comparing test performances at the level of the test items (e.g., comparing item difficulties), item characteristic curve theory, factor analysis, and a construct validity approach. An understanding of the psychological processes involved in performing the tasks involved in taking a test item, or performing an experimental task in a laboratory, is prerequisite to making judgments as to whether a test is measuring the same thing in two cultures. The methods described provide ways to improve the understanding of such processes. (RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED131093

RM-76-7

RESEARCH MEMORANDUM

HOW TO TELL IF A TEST MEASURES THE SAME THING IN DIFFERENT CULTURES

Norman Frederiksen

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented as the introduction to a symposium at the Third International Association for Cross-Cultural Psychology Congress, Tilburg, The Netherlands, July 13, 1976.

Educational Testing Service
Princeton, New Jersey
August 1976

TM005 761

HOW TO TELL IF A TEST MEASURES THE SAME THING IN DIFFERENT CULTURES

Eckensberger (1973) has defined cross-cultural research as "the explicit, systematic comparison of psychological measures obtained under different cultural conditions, in which cultural conditions...serve as the independent variables." Thus the hypothesis being tested in a typical cross-cultural study is that different cultures result in different forms of behavior. The problem we wish to consider has to do with the dependent variables: how can we tell whether or not a test (or other measurement procedure) measures the same psychological construct in different cultures? If it does not, the results of the investigation are likely to be erroneous. If in general tests do not measure the same thing in different cultures, or even if it is impossible to tell, cross-cultural psychology is in trouble.

That there may be a problem is well demonstrated by a study done by John French (1965) about ten years ago. He worked with subjects who would ordinarily be regarded as coming from one culture, and a rather homogeneous one at that. His 200 subjects were male Princeton University students, and Princeton High School seniors who planned to go to college. He administered to the students a battery of 15 tests--three tests to mark each of five cognitive factors. He later gave individually to the same subjects other forms of the same tests, and

the subjects were asked to "think aloud" as they took the tests and were questioned about how they went about solving the problems. For most of the tests he was able to identify different approaches, or problem-solving styles, and to form dichotomies of students on the basis of their methods of solving the problems. He then factor-analyzed the matrix of intercorrelations of the 15 tests separately for each subgroup of each dichotomy. In a number of instances he found quite different factor loadings and factor intercorrelations. The results showed that even the supposedly "pure-factor" tests used in the study did not measure the same things for all the subjects, even for the relatively homogeneous group of students in the sample.

For example, two subgroups were formed on the basis of whether an "analytic" or a "visualization" approach was used in solving Thurstone's Cubes test, a measure of spatial ability. For the subgroup using the visualization approach, the loading on a spatial-ability factor was .5, and for those using an analytic approach, the loading was nearly zero (.07). Apparently the use of analytic methods completely destroyed the capacity of the Cubes test to measure spatial ability. Other loadings showed that use of the analytic method transforms the test from a test of spatial ability to a test of inductive reasoning.

The term "analytic approach" suggests a problem-solving style like Witkin's field independence. French comments that many of the dichotomies he identified seem to be similar to such cognitive styles as focusing-scanning and field dependence-independence. Since cultural groups around the world are known to differ markedly with regard to

cognitive style variables, it appears that cross-cultural psychologists do have a problem. The concern is of course not necessarily limited to paper-and-pencil tests of cognitive abilities; the problem is potentially present whether the dependent variables are based on interviews, dream analysis, Rorschach protocols, or learning experiments; and it exists for any cultural comparison, whether it involves males and females, 8-year-olds and 12-year-olds, lower-class and middle-class Dutchmen, or Eskimos and Temne.

One reason the problem exists is that performance is influenced by many factors other than the amount of the construct that exists in the subject. Some children do poorly on perceptual tests because they are impulsive; some adults excel on reasoning tests because they possess certain mathematical skills; variations in social-desirability bias may influence the way students mark answers on a personality inventory; one school boy may do poorly as an athlete because he lacks competitiveness. Even Jimmy and Johnny, two pupils sitting beside each other in the same classroom, may, as we have seen, earn the same score on a test of cognitive ability by using quite different methods.

How to tell whether or not a test measures the same thing in two cultures is thus an important question for any psychologist and it is particularly important for cross-cultural psychologists. The purpose of this symposium is to examine a number of different ways of answering the question.

The variety of methods of approaching the problem is wide, ranging from some that are quite obvious and simple-minded to those that require

a considerable degree of statistical and psychological sophistication. The simpler methods include those used by John French in finding hypotheses on which to base his dichotomies--the "think aloud" technique and interviewing the candidates about how they solved the problems. It would also be possible to analyze the scratch work made by a candidate while he takes a math test, or to observe carefully the behavior of a subject in an experiment in an attempt to find cues as to how he performed the task. Such methods would be useful in providing hypotheses, but it would be risky to conclude on the basis of such evidence that two tests measure the same thing.

Another approach, which was dealt with quite fully in the Istanbul Conference (Cronbach and Drenth, 1972), involves doing things to make the groups more similar with regard to test-taking abilities and attitudes. Such methods would include use of pantomime or moving pictures to give instructions, especially when there are language problems; coaching and practice in taking test items; special incentives to control motivation and increase competitiveness; employment of people from the appropriate cultures as test makers and test administrators; insuring that the problems posed do not require information not provided in the culture; and so on. These methods may be expected to reduce the variance attributable to factors not related to the construct being measured, but it cannot be assumed that they will be completely successful. Positive evidence should be sought that the tests do indeed measure the same thing in the cultures being compared.

The opposite of this approach, in a sense, is to make the tests different in such a way that they measure the same construct--so that they are "functionally equivalent." For example, if it is found that members of an African tribe typically sort familiar objects into different categories than do, say, American high school students, a classification test might be modified in such a way that stimulus objects that are indigenous to each culture are used; or the scoring method might be changed so that for scoring purposes different sortings are judged to be equivalent. This approach probably has merit, although it seems risky. The method needs validation to at least the same extent as does the administration of identical tests to both groups.

There are a great variety of approaches to the problem that require use of statistical methods. One statistical method that immediately comes to mind is analysis of covariance, in which slopes and intercepts of regression lines for two or more cultural groups are compared. The method has been widely used in the United States in studying the fairness of tests for different ethnic groups. The method requires a criterion measure; the regressions of this criterion on the test in question are compared for two or more cultural groups. A major difficulty is that the criterion measure must be assumed to be unbiased, and often there is as much reason to question the fairness of the criterion as the test. Many variants of this kind of solution to the problem have been proposed (Cleary, Thorndike, Darlington, Nancy Cole, Novick,

etc.); but most of them are not really relevant to our question of whether the test is measuring the same thing. They deal with the social question of how tests should be used in decision making when it is desired to compensate for past disadvantages to a minority group while not being unjust to the majority group. The value judgments required to deal with social problems of test use are important but need not concern us here.

One statistical approach to the problem of how to judge whether a test measures the same thing in two cultures is to compare test performances at the level of the test items. The simplest method would be to compare item difficulties--for example, by making a plot of percent right for Group A against percent right for Group B. If the points on the graph fall in a narrow band extending diagonally upwards, the relative difficulties of the items are comparable for the two groups; if it is found that there are many items of different difficulty, use of the test for comparing the groups would be questionable.

Lord has pointed out, however, that the proportion of correct answers is not a satisfactory measure of item difficulty. (See his paper in these Proceedings.) Item characteristic curve theory provides a better method of comparing two groups with respect to their performance on individual test items. The item characteristic curve of an item for one group may be compared with the item characteristic curve of the same item for a second group; any difference in the curves indicates some kind of bias in the item. Differences may be indicative of differences in item difficulty or in item validity, or both. Significance tests are available.

Other statistical approaches based on scores on entire tests, rather than on items, are possible. One approach which has been rather widely used is factor analysis. This method was illustrated in my earlier remarks about John French's work. The idea is that if the interrelationships among a variety of tests are similar for two cultures, as revealed by similar patterns of factor loadings and factor intercorrelations, the tests collectively are probably measuring similar constructs.

If one is concerned about one particular test, it would be possible to include it as an "extension variable" in a factor analysis of a variety of other measures that for theoretical reasons would be expected to be related to the test. The loadings of the test in question on each factor can be calculated, while the test does not itself influence the factor structure. If the factor loadings for two cultures are similar, and if the tests producing the factor structure are appropriately chosen, the probability that the test is measuring the same construct in both cultures is increased. Use of maximum likelihood methods of factor analysis makes possible tests of goodness of fit to an hypothesis.

Such procedures are in fact methods for comparing construct validities of the test for two cultures. Construct validity is perhaps the most powerful idea for dealing with our problem, provided that the test battery includes not only measures to which we would expect our test to be related on theoretical grounds, but also measures where we would expect no relationships (for comparing discriminant validities). In addition, the tests should use a wide variety of procedures, such as observations of behavior, ratings, self-reports, biographical data, free response items, and

objectively scorable items. Statistical models and computer programs (Jöreskog, 1970; Bentler, 1976) now exist which make such analyses quite feasible.

The construct validity approach, if the variables are properly chosen, will tell us something about the psychological processes typically used by the members of a cultural group in dealing with the items presented in a test or an experimental situation. For example, the loadings of a test on cognitive factors in John French's study tell us that for one subgroup the process of solving Cubes test items involves inductive reasoning. There are other ways of investigating process that may get us slightly closer to some underlying psychobiological factors. If these processes appear to be the same for two cultural groups, we have additional evidence suggesting that the tests measure the same construct.

What I have in mind are the techniques used by experimental psychologists in attempting to learn about the processes involved in perception, learning, remembering, problem solving, and so on. These psychologists typically use the hypothetical components of information processing systems, and they perform ingenious experiments aimed, for example, at distinguishing processes underlying recognition as compared with recall, or inferring the charactistics of cognitive structures produced by different learning methods. Some of these psychologists have turned their attention to processes such as are involved in taking the traditional kinds of tests and variables often used by cross-cultural psychologists, such as tests of reading and verbal ability. The idea

of comparing cultural groups with regard to the processes underlying the taking of a test or being a subject in a learning experiment is attractive. Some of the psychologists working in this area are David Klahr, Earl Hunt, and James Greeno.

An understanding of the psychological processes involved in performing the tasks involved in taking a test item, or performing an experimental task in a laboratory, is prerequisite to making judgments as to whether a test is measuring the same thing in two cultures. The methods described above provide ways to improve our understanding of such processes.

References

Bentler, P. M. Multistructure statistical model applied to factor analysis. Multivariate Behavioral Research, 1976, 11, 3-25.

Cronbach, L. J., & Drenth, P. J. D. (Eds.). Mental Tests and Cultural Adaptation. The Hague: Mouton, 1972.

Eckensberger, L. H. Methodological issues of cross-cultural research in developmental psychology. In John R. Nesselroade and Hayne W. Reese (Eds.), Life-Span Developmental Psychology. New York: Academic Press, 1973.

French, J. W. The relationship of problem-solving styles to the factor composition of tests. Educational and Psychological Measurement, 1965, 25, 9-28.

Jöreskog, K. G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.